ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
DOI: 10.17559/TV-20140807111130

PREDICTION OF MAGNETIC SUSCEPTIBILITY CLASS OF SOIL USING DECISION TREES

Meltem Kurt, Nevcihan Duru, M. Mucella Canbay, H. Tarik Duru

Original scientific paper
Magnetic susceptibility (MS) is a dimensionless proportionality constant that indicates the degree of magnetization of a material in response to an applied magnetic field. In our study, the focus is to predict the magnetic susceptibility classification of the soil by using data mining algorithms. Magnetic susceptibility values depend on the composition, grain size of magnetic minerals and their source, such as lithogenic, pedogenic and anthropogenic origins. In this paper, we applied two data mining classification algorithms which are called ID3 and C4.5 for predicting MS class and the degree of pollution along the Izmir area in Turkey. By applying the algorithms, possible MS classes are obtained, according to the heavy metal concentration (Pb, Cu, Zn, Co, Cd, Ni) values related to MS. The aim of applying the algorithms is constructing the decision tree and the rules so as to obtain MS values. Thus, errors resulting from the change of ambient conditions and the measurement difficulties are eliminated. According to the rules, we reached 82 % accuracy condition and it is shown that test values and the measurement values are compatible with each other.

Keywords: *data mining; classification; heavy metal contamination; magnetic susceptibility*

Predviđanje magnetske osjetljivosti tla primjenom dijagrama za donošenje odluka

Izvorni znanstveni članak
Magnetska osjetljivost (MS) je konstanta nedimenzijske proporcionalnosti koja pokazuje stupanj magnetizacije materijala u magnetskom polja. U našem radu, cilj je predviđanje klasifikacije magnetske osjetljivosti tla primjenom algoritama za dubinsko istraživanje podataka (dobivanje korisnih, ranije nepoznatih podataka računarskom analizom velikih baza podataka). Vrijednosti magnetske osjetljivosti ovise o sastavu, veličini zrna magnetičnih minerala i njihovih izvora, litogeničnog, pedogeničnog i antropogeničnog porijekla. U radu smo primijenili dva d klasifikacijska algoritma za dubinsko istraživanje podataka nazvana ID3 i C4.5 za predviđanje vrste MS i stupnja zagađenja u području Izmir u Turskoj. Primjenom algoritama, dobivaju se moguće vrste MS prema vrijednostima koncentracije teških metala (Pb, Cu, Zn, Co, Cd, Ni). Cilj primjene algoritama je izrada dijagrama i pravila za donošenje odluka u svrhu dobivanja vrijednosti MS. Na taj način, eliminiraju se greške nastale promjenom uvjeta okoline i teškoća u mjerenju. Prema tim pravilima, dobili smo uvjete točnosti od 82 % i pokazali da su vrijednosti ispitivanja i vrijednosti mjerenja međusobno kompatibilne.

Ključne riječi: *dubinsko istraživanje podataka; klasifikacija; magnetska osjetljivost; zagađenost teških metala*

1 Introduction

The minerals that are present in soil are either natural (through lithogenesis, pedogenesis) or of anthropogenic origin (industrial residues). The magnetic mineral content of the soil can be expressed in very broad terms by its magnetic susceptibility [1]. Magnetic susceptibility is a measure of iron-bearing components in a material and it can be used to identify the type of the material on which the test is conducted as well as the amount of the iron-bearing minerals that the material contains [2]. Many studies are available in literature where the heavy metal contamination and industrial activities causing soil, air or water pollution were investigated [3 ÷ 7]. In addition, magnetic susceptibility was shown to be a highly useful indicator of industrial pollution, gas emission into air due to traffic and other atmospheric pollutants [1, 8 ÷ 20].

In the recent years, data mining studies found a place in the environmental geophysics publications. Especially in the evaluation of the results of field measurements these studies offer different interpretations to investigator. Studies about environmental geophysics have been moved to a different dimension with data mining methods.

In the literature some studies about data mining and environmental geophysics have been done. For example: Hanesch et al. [21] using fuzzy C-means cluster analysis and Non-Linear mapping techniques topsoil data from locations were analysed and link was observed between magnetic susceptibility and the heavy metal content with their method. Vibha et al. [22] presented an efficient hybrid model that was achieved by first clustering the data and then classifying it, and using the spatial conceptual information extracted from the environmental variables. Preetz et al. [23] introduced a classification system to assess soil magnetic susceptibilities from geoscientific maps. Canbay et al. [24] applied a data mining classification algorithm which is called C4.5 for predicting MS class and the degree of pollution along the Izmit area in Turkey. But we surveyed the literature on Magnetic susceptibility (MS) prediction with data mining methods and did not come across any study.

Pollution is a subject of current interest and there is a need for monitoring techniques developed by several fields of research, in order to analyse the distribution and the reach around the contamination sources. Although the man-made contribution of heavy metals and other pollutants can be studied by careful chemical methods (time-consuming, laborious and costly), magnetic monitoring constitutes an alternative tool for pollution studies. The relationship between both kinds of variables constitutes complex cases of non-linear mathematics. In consequence, multivariate techniques that have become necessary and used to investigate the problem, multivariate statistical analyses were investigated for magnetic monitoring in soils. Furthermore a classification and the need for prior knowledge also may be the case, long-term and sometimes the actual physical properties of soil samples can be lost to avoid taking this measurement in site. The soil data base comprised of pedological, geochemical and geological data and magnetic susceptibility data makes it necessary to evaluate the combination of multivariate study. We suggested using data mining techniques instead of these multivariate techniques.

The focus of our study is to predict the MS classification of the soil using data mining techniques.

MS values depend on the composition, grain size of magnetic minerals and their source, such as lithogenic, pedogenic and anthropogenic origins. In this paper, we applied two data mining classification algorithms which are called ID3 and C4.5 for predicting MS class and the degree of pollution along the Izmit area in Turkey. In our study, possible MS classes are obtained, according to the heavy metal concentration (Pb, Cu, Zn, Co, Cd, Ni) values related to MS. It is shown that test values and the measurement values are compatible with each other. The main aim of this paper is to determine the relationship between the data mining and heavy metal contamination via magnetic susceptibility measurement in the Kocaeli, Turkey area.

2 ID3 and C4.5 decision tree algorithms

Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas [25]. The developments of computer technology have created too much data on the other hand too little information. Therefore we need to extract useful information from the large chunk of data. The knowledge discovery process basically has seven steps: these are data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Steps 1 through 4 are different forms of data pre-processing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base [26].

In data mining, predictive and descriptive models can be separated as two main headers. In the predictive model, firstly a model is designed from the data that is known its result. After that this created model can predict future outcomes, results of which are unknown. In the descriptive model, the patterns that are extracted from existing data can be used to make a decision.

Data mining methods are grouped as classification, association rules and clustering. The classification method consists of decision trees, neural networks, Bayesian Network, Bayesian Classification.

In this study, we used decision tree method, ID3 [27] and C4.5 [28] algorithms, to classification of soil magnetic susceptibility.

ID3 algorithm selects the best attribute based on the concept of entropy and information gain for developing the tree. C4.5 algorithm acts similar to ID3 but improves a few of ID3 behaviours:

- A possibility to use continuous data.
- Using unknown (missing) values which have been marked by "?".
- Possibility to use attributes with different weights.
- Pruning the tree after being created [29].

In ID3 algorithm,
S: a set of s data samples,
m: number of values,
C_i: different classes $i = 1, 2,..., m$,
S_i: samples in C_i

$$I(S_1, S_2, ..., S_m) = -\sum_{i=1}^{m} p_i - \log_2(p_i). \quad (1)$$

To classify given data, required information amount is calculated with Eq. (1). p_i is the probability of an arbitrary data object belonging to C_i: $p_i = s_i/s$ [30].
$(a_1, a_2,..., a_v)$: A that is an attribute has v different values,
$(S_1, S_2,..., S_v)$: S is divided subset by using A,
S_j: occurs S and A has a_j value of data samples.

If A is selected as the test attribute, the entropy, required dividing sample set with A, will be calculated with Eq. (2).

$$E(A) = \sum_{i=1}^{v} \frac{s_{1j} + ... + s_{mj}}{s} I(s_{1j}, ..., s_{mj}) \quad (2)$$

Using attribute A on the current branch node the information gain, obtained by the corresponding sample collection divided, is calculated with Eq. (3).

$$Gain(A) = I(S_1, S_2, ..., S_m) - E(A). \quad (3)$$

Table 1 ID3 Algorithm pseudo code

```
Input values:
D: Dataset, A: Attributes, T: Target
ID3(D, A, T)
    Create node t,
    If all samples in D>0
        label (t) ="+",
        return (t),
    If all samples in D<0
label (t) ="−",
        return (t),
label (t)= class (D, T)
If A=0
        return (t),
    end if
let newA from A, //newA has maximum information gain
For each all possible value "a" in newA do
        Add new tree branch below t,
        Test newA= a,
Da is subset of D,
    If Da=0
        Create node newt,
        add leaf node with label (newt)= class (D, T),
createEdge (t, a, newt),
    else
        add sub tree ID3(Da, A \ {newA}, T),
    end if
end do
return (t)
```

ID3 algorithm pseudo code is given in Tab. 1.

In C4.5 algorithm used divide and conquer paradigm, based on multi-branched recursion, at the last result very high accuracy is procured.
In C4.5 algorithm,
T: datasets,
C: is collection as $(C_1, C_2, C_3,..., C_k)$,
v: is a property, it also divides T into subsets, has non-coincidence n value. $(v_1, v_2, v_3, ..., v_n)$,
$(T_1, T_2, T_3,..., T_n)$ that are values of all instances in T_i are v_i,
$|T|$: number of examples in T,

$|T_i|$: number of examples in $V = v_i$,
$|C_{jv}|$: number of examples with categories C_j in $V = v_i$,

$$P(C_j) = \frac{|C_j|}{|T|} = freq(C_j, T) \tag{4}$$

The occurrence probability of C_j can be calculated with Eq. (4).

$$P(v_i) = \frac{|T_i|}{|T|} \tag{5}$$

The occurrence probability of $V = v_i$ can be calculated with Eq. (5).

$$P(C_j \mid v_i) = \frac{|C_{jv}|}{|T_i|} \tag{6}$$

The conditional probability of the type of C_j in the cases of attribute $V = v_i$ is calculated with Eq. (6).

$$\begin{aligned} H(C) &= -\sum_j P(C_j) \cdot \log P(C_j) = \\ &= -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} \cdot \log \frac{freq(C_j, T)}{|T|} = Info(T). \end{aligned} \tag{7}$$

With Eq. (7) entropy of information can be calculated.

$$\begin{aligned} H\left(\frac{C}{V}\right) &= -\sum_j p(v_j) \cdot \sum_j p \cdot \left(\frac{C_j}{v_i}\right) = \\ &= -\sum_{i=1}^{n} \frac{|T_i|}{|T|} Info(T_i) = Info_v(T). \end{aligned} \tag{8}$$

Conditional entropy can be found with Eq. (8).

$$\begin{aligned} I(C,V) &= H(C) - H(C/V) \\ &= Info(T) - Info_v(T) = gain(v) \end{aligned} \tag{9}$$

Information gain can be calculated with Eq. (9).

$$\begin{aligned} H(V) &= -\sum_j p(v_i).\log p(v_i) \\ &= -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \log \sum_{i=1}^{n} \frac{|T_i|}{|T|} \\ &= split\_Info(v) \end{aligned} \tag{10}$$

The information entropy of attribute is found with Eq. (10) and finally with Eq. (11) gain ratio is calculated.

$$\begin{aligned} gain\_ratio(v) &= I(C,V) / H(V) = \\ gain(v) &/ split\_Info(v) \end{aligned} \tag{11}$$

C4.5 algorithm pseudo code is given in Tab. 2.

Table 2 C4.5 Algorithm pseudo code

```
Input values:
D: Dataset, T: Tree
T={};
    If D is pure or stopped
        break
    end if
    for all attribute a ϵ D do
        calculate information-theoretic criteria,
    end for,
a*=Best attribute,
T*=Create decision node  for finding a* in the root,
D*=Sub-datasets form D based on a*,
For all D* do
        T*=C4.function(D*),
        add T* to the corresponding branch of T,
 end for
return (T)
```

3 Magnetic susceptibility

Soil samples were collected vertically from a depth of 0 ÷ 30 cm at 13 stations situated on the 300 km^2 area with an average grid density of 10 km in the Kocaeli region. Stations were located at the middle Eocene aged Çaycuma formation and the rest of them were located at the Quaternary aged alluvium. The Çaycuma formation is composed of sandstone, claystone, marl, limestone and pebbles. Particle size varies between 0,0013 and 40 mm in this area. At last, samples were taken with plastic tubes at different depths within 30 cm of investigation depth at each station and then mixed as a composite sample for chemical analysis. Stations were chosen in a rural area around the roads and the others were selected close to site-specific pollution sources such as industrial plants near the main roads.

Susceptibility can be useful and very high sensitive and speed parameter of mineralogy and granulometry. Magnetic techniques have been applied by environmental scientists with demonstrable success in the pollution studies. Many anthropogenic emissions contain various particles, which cause heavy metal pollution of soils in industrial areas. Fundamentally, magnetic susceptibility can give a general view of the degree of pollution. Today, very often this method is used for agriculture.

Different authors studied the soil, air and water pollutions caused by heavy metal constructions and other industrial activities.

Magnetic susceptibility was shown to be highly useful in investigating industrial pollutants, traffic emission, and other atmospheric pollutants. The use of magnetic measurements as proxy of heavy metal pollution is based on the fact that origins of heavy metals and magnetic particles are genetically related. Environmental magnetism studies have demonstrated the relationship between heavy metal contents, and magnetic, lithological and pedological properties in soils. Several studies confirmed direct correlation between the magnetic susceptibility of contaminated soils and the presence of hydrocarbons and certain heavy metals (Pb, Zn etc.)

4 Materials and methods

Magnetic susceptibility measurements were collected from 13 different stations and different environmental

settings: a heavy industrial area with main roads of heavy traffic, and a rural area around the roads. 93 samples were taken vertically from 3 different layers (5, 10 and 15 cm) in 13 stations. The surface measurements were performed using an SM-20 and MS-2 Bartington loop sensor with a diameter of 185 mm at the stations. The penetration depth is about 30 cm, after the magnetic susceptibility measurement in laboratory, heavy metal (Pb, Cu, Zn, Co, Cd and Ni) contents and concentrations of the samples were determined using 6001 model Atomic Absorption Spectrometer of Shımadzu. Samples were taken with plastic tubes at different depths within 30 cm of investigation depth at each station and then mixed as a composite sample for chemical analysis. 1,0 ± 0,09 g soil samples were weighed and placed in platinum or porcelain crucibles. In the ash furnace, the temperature was gradually increased to 900 °C. The samples were then left to cool in the furnace and then taken into 100 ml beakers, in which 10 ml HNO3 and 30 ml HCl (both acids should be concentrated, 'king water') were added. The mixture was dried by evaporating the liquid mixture in a fume cupboard and then 5 ml of concentrated HCl was added to the mixture after which it was dried by evaporating the mixture. The remaining mixture was dissolved in a small amount of HCl just enough to dissolve the samples. The volume was finally brought up to 250 ml with HCl solution (5 %).

Magnetic susceptibility values measured in field (topsoil magnetic susceptibility measurements), mass-specific magnetic susceptibility values measured in laboratory and heavy metal concentrations are given in Tab. 3.

Table 3 Some examples of heavy metal concentration (mg/kg), topsoil magnetic susceptibility and mass-specific magnetic susceptibility values

Heavy metal	Pb	Cu	Zn	Co	Cd	Ni	Topsoil magnetic susceptibility field measurements	Mass-specific magnetic susceptibility lab. measurements
Mc 1 ÷ 5 cm	214,13	76,48	40,2	4,75	0,07	3,35	15,4	135,294
10 cm	53,4	34,13	15,18	3,335	0,05	5,22	19,1	102,353
15 cm	132,3	65,55	69,53	8,75	0,08	2,13	15	138,824
Mc 2 ÷ 5 cm	48,1	54,9	64,73	2,18	0,06	2,3	13,7	7,333
10 cm	7,5	718,78	90,6	39,38	0,03	4,5	13,6	8,333
15 cm	39,28	147,88	54,7	42,98	0,06	28,28	14,6	11,333
Mc 3 ÷ 5 cm	39,28	53,3	80,35	17,03	0,04	1,8	16,5	118,159
10 cm	7,5	49,58	95,65	7,6	0,07	2,85	17,1	65,985
15 cm	58,7	61,28	74,25	25,88	0,07	5,03	16,9	73,657
Mc 4 ÷ 5 cm	12,8	59,15	91	21,58	0,06	5,32	85,7	32,344
10 cm	32,23	76,48	55,03	16,45	0,06	4,65	93,1	4,688
15 cm	41,05	65,28	95,1	10,18	0,05	2,56	75,6	2,578
Mc 5 ÷ 5 cm	49,88	71,68	69,13	40,7	0,07	11,44	72,9	50,885
10 cm	7,5	86,85	99,1	5,38	0,1	3,28	77,7	112,832
15 cm	55,18	86,6	68,08	29,58	0,08	5,17	53,2	108,407
Mc 6 ÷ 5 cm	41,05	253,63	86,85	61,83	0,08	31,54	55,8	37,886
10 cm	48,1	79,68	83,23	18,45	0,09	3,46	54,2	6,167
15 cm	62,25	160,65	60,95	37,55	0,07	18,54	54,8	6,167
Mc 7 ÷ 5 cm	21,63	60,23	29,2	17,03	0,01	5,43	49,7	67,241
10 cm	49,88	58,9	37,95	10,45	0,08	2,79	50,1	15,517
15 cm	53,4	74,35	64,08	21,03	0,07	3,78	12,6	11,379
Mc 8 ÷ 5 cm	37,53	70,35	66,8	20,15	0,08	2,73	15,9	92,632
10 cm	101,1	49,03	89,88	21,88	0,07	3,96	15,9	17,895
15 cm	78,13	111,1	68,33	33	0,08	14,26	15,3	41,053
Mc 9 ÷ 5 cm	58,7	51,18	89,08	12,18	0,09	3,26	13,9	75,836
10 cm	74,6	26,4	40,75	8,18	0,05	3,54	12,2	86,989
15 cm	65,78	64,48	36,75	27,57	0,08	4,37	11	79,182
Mc 10 ÷ 5 cm	118,75	37,05	27,05	21,03	0,08	4,08	11,9	251,111
10 cm	79,9	30,13	16,15	18,72	0,08	6	20,1	304,444
15 cm	101,1	42,9	33,05	33	0,04	4,8	21,6	53,333

5 Improved method

In this study we improved a method that can be classify of soil magnetic susceptibility with classification algorithms also predict results of new data. We presented prediction success rate for each algorithm. There are Pb, Cu, Zn, Co, Cd and Ni heavy metal measurements of soil samples in our dataset. The reference values for heavy metal concentrations are given in Tab. 4 [31].

The classification of soil magnetic susceptibility results that is based on thresholds given in Tab. 5, were determined by experimental measurements in the city of Kocaeli.

Table 4 Reference heavy metal concentrations acquired from the literature (mg/kg)

	Cd	Co	Cr	Cu	Ni	Pb	Zn
Average for the Earth crust	0,16	-	-	68	99	13	76
Average for the Earth soil	0,06	8	100	30	40	10	50
Average for the Turkish soil	1,00	20	100	50	30	50	150

Table 5 The classification of soil magnetic susceptibility

Classification (ID3)	Classification (C4.5)	Magnetic susceptibility $\kappa / 10^{-5}$ SI
few	1	0 ÷ 30
medium	2	30 ÷ 60
high	3	>60

The Çaycuma formation is composed of sandstone, claystone, marl, limestone and pebbles. Generally particle size varies between 0,0013 and 40 mm in this area. The sampled region has very small grain size of magnetic minerals and high penetration capacity. Soil example is taken as mixed. We also took large number of samples from different featured rock samples and thus improved method can extract more general rules about classification of magnetic susceptibility of these soil examples.

In Tab. 6, some examples of train dataset are given dataset.xlsx file.

Table 6 Some examples of dataset.xlsx file

Id	Pb	Cu	Zn	Co	Cd	Ni	Results for	
							C4.5	ID3
1	12,8	59,15	91	21,58	0,06	5,32	3	high
2	32,23	76,48	55,03	16,45	0,06	4,65	3	high
3	55,18	86,6	68,08	29,58	0,08	5,17	2	medium
4	21,63	60,23	29,2	17,03	0,01	5,43	2	medium
5	53,4	74,35	64,08	21,03	0,07	3,78	1	few
6	37,53	70,35	66,8	20,15	0,08	2,73	1	few
7	101,1	49,03	89,88	21,88	0,07	3,96	1	few

Heavy metal measurements are used as an attribute. "results for C4.5"column that is the attribute of the target class in the dataset only used to compose C4.5 decision tree, with the same method using "results for ID3" attribute column ID3 algorithm's decision tree can be drawn. We used all of the data in this dataset to extract classification rules.

According to our method, both algorithms ID3 and C4.5 are used. Basically a dataset can have numeric or categorical data. ID3 algorithm clusters only categorical data however C4.5 algorithm clusters only numeric data. Our method can cluster both data types. Converting numerical value to categorical value can be done automatically or user can divide category of values between minimum and maximum data value. In an automatical option, minimum and maximum attribute values are calculated then these two values are subtracted from each other. Result value is divided category number thus categorical data is created for each attribute. We used all measurement values in dataset.xlsx file for classification. All data in our dataset is numerical value for that reason we converted every heavy metal attribute's value to categorized value. In this method, we calculated minimum and maximum category number for each attribute. These values will reduce or increase depending on the size of data in our dataset. After we calculated minimum and maximum category value for each attribute, we decided to use eight categories to extract more rules. All of the category values calculated with using automatical option. Finally we extracted a set of rules using ID3 algorithm. These rules are given in Tab. 7. According to our dataset Cd is founded as decision point. Namely if you use ID3 algorithm in our method, firstly Cd heavy metal will be controlled then other heavy metal values will be effective.

Table 7 The extracting rules from ID3 algorithm

0,06625< Cd < 0,0775 and Pb> 188,30125 →RESULTSFORID3 = few.
0,06625< Cd < 0,0775 and Pb< 33,32875 →RESULTSFORID3 = few.
0,06625< Cd < 0,0775 and 33,32875 <Pb< 59,1575 and 24,54875 < Co < 32,005 →RESULTSFORID3 = few.
0,06625 < Cd < 0,0775 and 33,32875 <Pb< 59,1575 and 39,46125 < Co < 46,9175 →RESULTSFORID3 = high.
0,06625 < Cd < 0,0775 and 33,32875 <Pb< 59,1575 and 17,0925 < Co < 24,54875 →RESULTSFORID3 = few.
0,06625< Cd < 0,0775 and 59,1575 <Pb< 84,98625 →RESULTSFORID3 = medium.
0,06625< Cd < 0,0775 and 84,98625 <Pb< 110,815 →RESULTSFORID3 = few.
0,06625< Cd < 0,0775 and 110,815 <Pb< 136,64375 →RESULTSFORID3 = few.
0,04375< Cd < 0,055 and Zn < 25,65 →RESULTSFORID3 = few.
0,04375< Cd < 0,055 and 85,9 < Zn < 97,95 →RESULTSFORID3 = high.
0,04375< Cd < 0,055 and 37,7 < Zn < 80,002 →RESULTSFORID3 = few.
0,0775< Cd < 0,08875 and Co < 9,63625 →RESULTSFORID3 = few.
0,0775< Cd < 0,08875 and 24,54875 < Co < 32,005 and 23,3985 <Pb< 38,955 →RESULTSFORID3 = high.
0,0775< Cd < 0,08875 and 24,54875 < Co < 32,005 and 38,955 <Pb< 59,1575 →RESULTSFORID3 = medium.
0,0775< Cd < 0,08875 and 24,54875 < Co < 32,005 and 59,1575 <Pb< 84,98625 →RESULTSFORID3 = few.
0,0775< Cd < 0,08875 and 24,54875 < Co < 32,005 and 136,64375 <Pb< 162,4725 →RESULTSFORID3 = few.
0,0775< Cd < 0,08875 and Co>54,37375 →RESULTSFORID3 = medium.
0,0775< Cd < 0,08875 and 9,63625 < Co < 17,0925 →RESULTSFORID3 = medium.
0,0775< Cd < 0,08875 and 17,0925 < Co < 24,54875 →RESULTSFORID3 = few.
0,0775< Cd < 0,08875 and 32,005 < Co < 39,46125 →RESULTSFORID3 = few.
0,0775< Cd < 0,08875 and 39,46125 < Co < 46,9175 →RESULTSFORID3 = few.
0,055 < Cd< 0,06625 and 33,32875 <Pb< 59,1575 →RESULTSFORID3 = few.
0,055 < Cd < 0,06625 and Pb< 33,32875 →RESULTSFORID3 = high.
0,055 < Cd < 0,06625 and 84,98625 <Pb< 110,815 →RESULTSFORID3 = few.
0,055 < Cd< 0,06625 and 110,815 <Pb< 136,64375 →RESULTSFORID3 = few.
0,02125< Cd < 0,0325 →RESULTSFORID3 = few.
0,0325< Cd < 0,04375 →RESULTSFORID3 = few.
Cd>0,08875 and Zn>97,95 →RESULTSFORID3 = high.
Cd>0,08875 and 73,85 < Zn < 85,9 →RESULTSFORID3 = medium.
Cd>0,08875 and 85,9 < Zn < 97,95 →RESULTSFORID3 = few.
Cd < 0,02125 →RESULTSFORID3 = medium.

We also got new rules from C4.5 algorithm for the same dataset using improved method. These new rules are given in Tab. 8. Using C4.5 algorithm in our dataset Pb is founded as decision point. As will be understood from

Tab. 8 Pb heavy metal is the highest possible decisive criterion for classification result.

Tab. 8 The extracting rules from C4.5 algorithm

Pb< 62,25 and Cu < = 54.9 →RESULTSFORC4.5 = 1.
Pb< 62,25 and Cu > 54.9 and Zn < = 90,6 and Cd < = 0,07 and Zn < = 61, 03 →RESULTSFORC4.5 = 2.
Pb< = 62,25 and Cu > 54,9 and Zn < = 90,6 and Cd > 0,07 →RESULTSFORC4.5 = 2.
Pb< = 62,25 and Cu > 54,9 and Zn < = 90,6 and Cd < = 0, 07 and Zn > 61, 03 →RESULTSFORC4.5 = 1.
Pb< = 62,25 and Cu > 54,9 and Zn > 90, 6 →RESULTSFORC4.5 = 3.
Pb> 62,25 RESULTSFORC4.5 = 1.

Basically both tables Tab. 7 and Tab. 8 refer the same results with different classification name because of different algorithms. Namely "1" and "few", "2" and "medium", "3" and "high" definitions have the same meaning that can be shown in Tab. 5. These rules can be used for the new test datasets to get classification results. Thanks to improved algorithm, we can also draw individual decision trees from ID3 and C4.5 algorithms' rules. In Fig. 1 C4.5 algorithm's decision tree is given but we did not show ID3 algorithm's decision tree because it is too big.

According to Fig. 1 heavy metal concentrations of Pb, Cu, Zn, Cd are important to predict MS value with C4.5 algorithm. The most important metal is Pb then we look at Cu concentration and then sequentially Zn, Cd, at last looking back to Zn concentration. The improved method also finds classification result of new data. If you enter new values for each attribute, the algorithm easily classifies this new data and returns classification result. We used soil samples from different sampling stations for testing. Their average heavy metal concentrations, topsoil MS field measurements and mass-specific MS laboratory measurements are given in Table 9 [31].

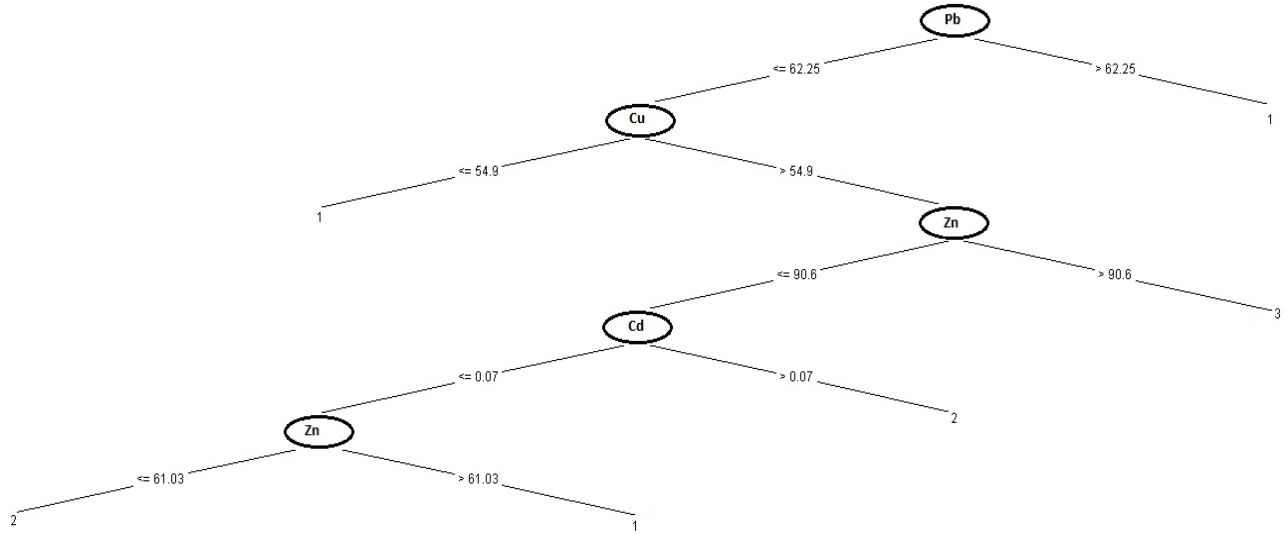

Figure 1 C4.5 algorithm'sdecision tree

Table 9 Average heavy metal concentrations and the on-site and in the MS measurements for each sampling station

Sampling stations	Average heavy metal concentrations						Topsoil MS field measurements × 10^{-5} (SI)	Mass-specific MS lab. measurements × 10^{-5} (SI)
	Pb (mg/kg)	Cu (mg/kg)	Zn (mg/kg)	Co (mg/kg)	Cd (mg/kg)	Ni (mg/kg)		
Mc 1	133,20	58,70	41,60	5,60	0,06	3,50	16,5	125,4
Mc 2	31,60	307,10	70,00	28,10	0,05	15,20	13,9	8,9
Mc 3	35,10	54,70	83,40	16,80	0,06	3,20	16,8	85,9
Mc 4	28,60	66,90	80,30	16,00	0,05	4,10	84,8	13,2
Mc 5	37,50	81,70	78,70	25,20	0,08	6,60	67,9	90,7
Mc 6	50,40	164,60	77,00	39,20	0,08	17,80	54,9	16,7
Mc 7	41,60	64,40	43,70	16,10	0,05	4,00	37,4	31,3
Mc 8	72,20	76,80	75,00	25,00	0,07	6,90	15,7	50,5
Mc 9	66,30	47,30	55,50	15,90	0,07	3,70	12,3	80,6
Mc 10	99,90	36,60	25,40	24,20	0,06	4,90	17,8	202,9
Mc11	87,50	38,30	41,50	26,00	0,05	4,50	15,4	41,7
Mc 12	105,20	47,80	69,50	38,50	0,07	3,60	15,8	27,5
Mc 13	157,70	35,50	23,10	36,20	0,08	3,60	14,2	2,3

We took into account "topsoil MS field measurements" column in Tab. 9. These 13 different samples are used as test data over extracted rules in Table 7 and Tab. 8. Using improved method we aimed to achieve the same values with that column's classification value in accordance with Tab. 5. In Tab. 10 extracting results from ID3 and C4.5 algorithms and real topsoil MS field measurements given in Tab. 9 were compared with each other. That can be seen in Tab. 10, C4.5 algorithm's correct prediction rate is 12 % better than ID3s for these samples.

Additionally we applied extracted rules on 150 new / different soil measurements. According to our observations C4.5 algorithm's prediction rate is more successful than ID3 algorithms by about 43 %. As a result in this method magnetic susceptibility class can be estimated more correctly by using C4.5 algorithm and totally we reached 82 % accuracy condition on soil MS values.

Table 10 Test Results received from improved method

Test Data		For ID3 Algorithm			For C4.5 Algorithm		
Sampling stations	Topsoil MS field measurements	Classification topsoil MS field measurements values according to Tab. 4	Estimated class using improved method	Accuracy condition (true / false)	Classification topsoil MS field measurements values according to Tab. 4	Estimated class using improved method	Accuracy condition (true / false)
Mc1	16,5	few	few	true	1	1	true
Mc2	13,9	few	few	true	1	1	true
Mc3	16,8	few	few	true	1	1	true
Mc4	84,8	high	high	true	3	1	false
Mc5	67,9	high	high	true	3	2	false
Mc6	54,9	medium	medium	true	2	2	true
Mc7	37,4	medium	few	false	2	2	true
Mc8	15,7	few	medium	false	1	1	true
Mc9	12,3	few	medium	false	1	1	true
Mc10	17,8	few	few	true	1	1	true
Mc11	15,4	few	few	true	1	1	true
Mc12	15,8	few	few	true	1	1	true
Mc13	14,2	few	few	true	1	1	true
Rate of true prediction				10/13 (77 %)			11/13 (85 %)

6 Conclusion

Element variety created by the soil structure changing for sampling locations also changes the magnetic susceptibility. Another important result realized in the measurement stage was that the magnetic susceptibility decreases in the samples with high content variety. In addition, it can be said that when extreme minimum and maximum values are encountered in magnetic sensitivity in the short distant measurement locations where the kind of rocks and pollution source do not show variation. Another frequent situation was that changing element properties also changes anisotropic properties.

After the measurements in the fields, samples were taken to the laboratory at the same time, thus completing the measurement process as far as possible without samples losing their properties. The characteristics of natural environment cannot be protected so that digital difference between measurement values occurs.

In this study, chemical analyses results and field measurements were considered and then some rules were extracted. Thanks to these rules new heavy metal values' (concentration quantities) field measurement class are predicted. Field measurements namely topsoil MS field measurements are used as a target class for our dataset. ID3 and C4.5 classification algorithms have been applied for prediction of magnetic susceptibility of the soil. According to the heavy metal concentration and topsoil magnetic susceptibility values, the improved method can construct the decision tree and the rules and then predict the MS class of the new soil example.

As mentioned previously, some rules given in Tab. 7 and Tab. 8 are extracted. These rules will change when new/different dataset is used. Cd is the most efficient heavy metal over classification result according to ID3 algorithm, however Pb is the most efficient heavy metal in C4.5 algorithm. This difference comes from the mathematical calculations in these algorithms.

In addition to our study, for thousands of soil examples the decision tree may be more stable. The soil examples belong to one region. If we could use more examples from different regions of the country we could obtain more general rules. Using these rules, magnetic susceptibility measurements can be classified entering measured heavy metal values. Moreover it is not necessary to bring soil samples to laboratory environment to measure magnetic susceptibility. Thus, errors resulting from the change of ambient conditions are eliminated.

As a future work, other classification methods for example Random Forest algorithm or Naive Bayes can be applied. In addition, the methods can be compared on the same MS values for deciding the best algorithm. Also the algorithms should be applied to thousands of heavy metal values.

7 References

[1] Thampson, R.; Oldfield, F. Environment Magnetism. Effect of climate on the magnetic susceptibility of soils. Allen and Unwin, 1986.

[2] Canbay, M. Investigation of the Relation between Heavy Metal Contamination of Soil and Its Magnetic Susceptibility. // International Journal of Physical Sciences. 5, (5)2010, pp. 393-400.

[3] Le Borgne, E. Susceptibilité magnétique anormal du soil superficial. // Ann. Geophys. 11, (1955), pp. 399-419.

[4] Le Borgne, E. Influence du feu sur les propriétés magnétiques du sol et sur celles du schiste et du granite. // Ann. Geophys. 16, (1960), pp. 159-195.

[5] Vadiunina, A. F.; Babanin, V. F. Magnetic susceptibility of some soils in the U.S.S.R. // Soviet. Soil Sci. 6, (1972), pp. 106-110.

[6] Mullins, C. E.; Tite, M. S. Magnetic viscosity, quadrature susceptibility and frequency dependence of susceptibility in single–domain assemblages of magnetite and maghaemite.

// J. Geophys. Res. 78, (1973), pp. 804-809. DOI: 10.1029/JB078i005p00804
[7] Mullins, C. E. Magnetic susceptibility of the soil and its significance in soil science: a review. // J. Soil Sci. 28, (1977), pp. 223-246. DOI: 10.1111/j.1365-2389.1977.tb02232.x
[8] Hay, K. L.; Dearing, J. A.; Baban, S. M. J.; Loveland, P. A. Preliminary attempt to identify atmospherically-derived pollution particles in English topsoils from magnetic susceptibility measurements. // Physics and Chemistry of the Earth. 22, (1997), pp. 207-210. DOI: 10.1016/S0079-1946(97)00104-3
[9] Strzyszcz, Z.; Magiera, T. Magnetic susceptibility and heavy metals contamination in soils of Southern Poland. // Phys. Chem. Earth. 23, 9-10(1998), pp. 1127-1131.
[10] Durza, O. Heavy metals contamination and magnetic susceptibility in soils around metallurgical plan. // Phys. Chem. Earth, Part A solid Earth Geod. 24, 6(1999), pp. 541-543. DOI: 10.1016/S1464-1895(99)00069-1
[11] Kapička, A.; Petrovský, E.; Jordanova, N. Comparison Of In Sıtu Field Measurements of Soil Magnetıc Susceptibility with Laboratory Data. // Studiageoph. etgeod, (1997), pp. 41.
[12] Kapička, A.; Petrovský, E.; Ustjak, S.; Macháčková, K. Proxy mapping of fly-ash pollution of soils around a coal-burning power plant: a case study in the Czech Republic. // J. Geochem. Explor. 66, (1999), pp. 291-297. DOI: 10.1016/S0375-6742(99)00008-4
[13] Kapička, A.; Jordanova, N.; Petrovský, E.; Podrazský, V. Magnetic study of weakly contaminated forest soils. // Water Air Soil Pollut. 148, (2003), pp. 31-44. DOI: 10.1023/A:1025429928763
[14] Lecoanet, H.; Lévêque, F.; Seguna, S. Magnetic susceptibility in environmental applications: comparison of field probes. // Phys. Earth Planet. Inter. 115, (1999), pp. 191-204. DOI: 10.1016/S0031-9201(99)00066-7
[15] Lecoanet, H.; Lévêque, F.; Ambrosi, J. Magnetic properties of saltmarsh soils contaminated by iron industry emissions (southeast France). // J. Appl. Geophys. 48, (2001), pp. 67-81. DOI: 10.1016/S0926-9851(01)00080-5
[16] Knab, M.; Appel, E.; Hoffmann, V. Separation of the anthropogenic portion of heavy metal contents along a highway by means of magnetic susceptibility and fuzzy c-means cluster analysis. // Eur. J. Environ. Eng. Geophys. 6, (2001), pp. 125-140.
[17] Hanesch, M.; Scholger, R. Mapping of heavy metal loadings in soils by means of magnetic susceptibility measurements. // Environ. Geol. 42, (2002), pp. 857-870. DOI: 10.1007/s00254-002-0604-1
[18] Hanesch M, Scholger R, Rey D Mapping dust distribution around an industrial site by measuring magnetic parameters of tree leaves. // Atmos. Environ. 37, (2003), pp. 5125-5133. DOI: 10.1016/j.atmosenv.2003.07.013
[19] Hanesch, M.; Scholger, R. The influence of soil type on the magnetic susceptibility measured throughout soil profiles. // Geophys. J. Int. 161, (2005), pp. 50-56. DOI: 10.1111/j.1365-246X.2005.02577.x
[20] Lu, S. G.; Bai, S. Q.; Xue, Q. F. Magnetic properties as indicators of heavy metals pollution in urban topsoils: a case study from the city of Luoyang, China. // Geophys. J. Int. 171, (2007), pp. 568-580. DOI: 10.1111/j.1365-246X.2007.03545.x
[21] Hanesch, M.; Scholger, R.; Dekkers, M. J. The Application of Fuzzy C-Means Cluster Analysis and Non-Linear Mapping to a Soil Data Set for the Detection of Polluted Sites. // Phys. Chem. Earth. 26, (2001), pp. 885-891. DOI: 10.1016/S1464-1895(01)00137-5
[22] Vibha, L.; HarshaVardhan, G. M.; Prashanth, S. J.; Deepa-Shenoy, P.; Venugopal, K. R.; Patnaik, L. M. A Hybrid Clustering and Classification Technique for Soil Data Mining. // International Conference on Information and Communication Technology in Electrical Sciences, India, 2007, pp. 1090-1095.
[23] Preetz, H.; Altfelder, S.; Hennings, V.; Igel, J. Classification of Soil Magnetic Susceptibility and Prediction of Metal Detector Performance-Case Study of Angola. // Detection and Sensing of Mines. 730313, (2009), DOI: 10.1117/12.819394
[24] Canbay, M. M.; Duru, N.; Duru, H. T. Data mining application on Magnetic Susceptibility of the soil. // International Earth Science Colloquium on Aegean Region (IESCA). (2012), pp. 312.
[25] Hand, D. J. Data Mining: Statistics and More? // The American Statistician. 52, 2(1998), pp. 112-118.
[26] Han, J.; Kamber, M.; Pei, J. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann Publishers, 2012. DOI: 10.1007/978-1-4419-1428-6_3752
[27] Quinlan, J. R. Induction of Decision Trees. // Machine Learning. 1, 1(1986), pp. 81-106. DOI: 10.1007/BF00116251
[28] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
[29] Khoonsari, P. E.; Motie, A. A Comparison of Efficiency and Robustness of ID3 and C4.5 Algorithms Using Dynamic Test and Training Datasets. // International Journal of Machine Learning and Computing. 2, 5(2012), pp. 540-543.
[30] He, L.; Liying, C.; Guifen, C.; Dexin, L. Delineating Soil Nutrient Management Zones Based on ID Algorithm. // Proceedings of the Int. conference on Mechatronic Science, Electric Engineering and Computer, Jilin, China, 2011, pp. 1155-1159.
[31] Canbay, M. Investigation of the Relation between Heavy Metal Contamination of Soil and Its Magnetic Susceptibility. // International Journal of Physical Sciences, 5, 5(2010), pp. 393-400.

Authors' addresses

Meltem Kurt Pehlivanoglu, Research Assistant
University of Kocaeli,
Department of Computer Engineering,
Faculty of Engineering,
Umuttepe Campus, 41380 Izmit Kocaeli, Turkey
E-mail: meltem.kurt@kocaeli.edu.tr

Nevcihan Duru, Prof. Dr.
University of Kocaeli,
Department of Computer Engineering,
Faculty of Engineering,
Umuttepe Campus, 41380 Izmit Kocaeli, Turkey
E-mail: nduru@kocaeli.edu.tr

M. Mucella Canbay, Assoc. Prof. Dr.
University of Kocaeli,
Department of Geophysics,
Faculty of Engineering,
Umuttepe Campus, 41380 Izmit Kocaeli, Turkey
E-mail: mucella@kocaeli.edu.tr

H. Tarik Duru, Prof. Dr.
University of Kocaeli,
Department of Electrical Engineering,
Faculty of Engineering,
Umuttepe Campus, 41380 Izmit Kocaeli, Turkey
E-mail: tduru@kocaeli.edu.tr